DOT/FAA/AR-98/31

William J. Hughes Technical Center
Atlantic City International Airport
New Jersey, 08405

Development and Validation Plan for a Screener Readiness Test

PB98-168727

Eric C. Neiderman, Ph.D.
J. L. Fobes, Ph.D.

Aviation Security Human Factors
Program, AAR-510
William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

August 1998

This document is available to the public through the
National Technical Information Service (NTIS),
Springfield, Virginia 22161.

U.S. Department of Transportation
Federal Aviation Administration

REPRODUCED BY: NTIS
U.S. Department of Commerce
National Technical Information Service
Springfield, Virginia 22161

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.

Technical Report Documentation Page

1. Report No.
DOT/FAA/AR-98/31

PB98-168727

3. Recipient's Catalog No.

4. Title and Subtitle

DEVELOPMENT AND VALIDATION PLAN FOR A SCREENER READINESS TEST

5. Report Date

August 1998

6. Performing Organization Code
AAR-510

7. Author(s)
Eric C. Neiderman, Ph.D. & J. L. Fobes, Ph.D.

8. Performing Organization Report No.

9. Performing Organization Name and Address
U.S. Department of Transportation
Federal Aviation Administration Wm. J. Hughes Technical Center
Aviation Security Research and Development Division, AAR-500
Atlantic City International Airport, NJ 08405

10. Work Unit No.

11. Contract or Grant No.
DTFA03-98-D-00010

12. Sponsoring Agency Name and Address
U.S. Department of Transportation
Federal Aviation Administration
Associate Administrator for Civil Aviation Security, ACS-1
800 Independence Avenue, S.W.
Washington D.C. 20590

13. Type of Report and Period Covered

Project Plan

14. Sponsoring Agency Code
ACS-1

15. Supplementary Notes: Draft Prepared By:

William Maguire, Ph.D.
Veridian, Veda Operations
780 Falcon Circle, Suite 100
Warminster, PA 18974

Federal Data Corporation
Science and Engineering Division
500 Scarborough Drive
Egg Harbor Township, New Jersey 08234

16. Abstract This document describes a plan to develop and validate a reliable, non-biased, secure test for initial screener training which can be used as a measure of knowledge of the screener's role in threat detection and checkpoint operations before on-the-job training.

17. Key Words Initial Screener Training,
Knowledge, Skills, and Abilities (KSA)
Screener Readiness Test (SRT)

18. Distribution Statement This document is available to the public through the National Technical Information Service (NTIS), Springfield, Virginia 22161.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 17	22. Price

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ACSSP	Air Carrier Standard Security Program
ATA	Air Transport Association
CBT	Computer-Based Training
DCA	Ronald Reagan Washington National Airport
DIF	Differential Item Functioning
FAA	Federal Aviation Administration
IED	Improvised Explosive Device
KSA	Knowledge, Skills, and Abilities
OJT	On-the-job Training
QA	Quality Assurance
SME	Subject Matter Expert
SQA	Software Quality Assurance
SRT	Screener Readiness Test

1.0 INTRODUCTION

1.1 Background

The effectiveness of the national civil aviation security system is highly dependent upon people, especially those employed as checkpoint screeners. The training of these individuals is critical to their performance on the job. Therefore, the Federal Aviation Administration (FAA) is very interested in enhancing screener training and further improving their readiness for the job.

According to FAR § 108.17 (use of X-ray systems), air carriers are required to have a program for initial and recurrent training of operators of X-ray systems, which includes training in radiation safety, the efficient use of X-ray systems, and the identification of weapons and other dangerous articles. Section XIII of the Air Carrier Standard Security Program (ACSSP) presents the standards for training and testing of persons performing screening and security functions.

For many years, the only FAA-approved training was developed by the Air Transport Association (ATA). Completion of this 12-hour initial screener training program is based on passing an exam with 40 multiple choice questions and 40 X-ray images to assess mastery prior to on-the-job training (OJT). In April 1997, the FAA also approved the use of Safe Passage's Computer-Based Training (CBT) system for initial screener training prior to OJT. The CBT system has a library of test questions and the trainee is presented with unit tests, a 50-item content mastery final exam, and a 50-item threat image interpretation final exam to assess mastery.

The variety of training options is growing and the FAA needs a single uniform measure of mastery of the classroom knowledge necessary before a screener can graduate to OJT. As additional training systems are offered for initial screener training, each is expected to have a different test to assess mastery prior to OJT and screener certification. To address this issue, the FAA may field a test for screener readiness to enter the OJT phase of preparation. This document describes a plan to develop and validate a reliable, non-biased, secure test for initial screener training.

1.2 Project Goals

The goal is to develop a valid test for initial screener training that can run on both MacIntosh and PC platforms. This test will have the following characteristics:

- Self administered with minimal demands for oversight.
- Job-relevant, performance-oriented testing of screener content mastery.
- Test questions that accurately sample the entire range of screener job functions and roles.
- A large pool of test questions to minimize the likelihood of cheating and of the test being compromised.
- Automatic collection and secure storage of test scores/records.
- Good psychometric qualities including demonstrated criterion-related validity, test-retest reliability, and sub-modules that are internally consistent (i.e., inter-item reliability).
- Fair and unbiased against specific population groups (i.e., no adverse impact as defined by the Equal Employment Opportunity Act under Title VII of the Civil Rights Act of 1964).

- Upgradable with minimal effort to reflect changes in aviation security and the ACSSP.

2.0 MAJOR PROGRAM ACTIVITIES

2.1 Quality Assurance

Quality Assurance (QA) is the overall process of evaluations, inspections and audits conducted during the test's developmental process and its products to ensure that: 1) the process and products conform to their established plans and standards; 2) the final product(s) completely and accurately implements the system's functional, performance and operational requirements; and, 3) the test is built to the highest quality attributes possible (reliability, maintainability, supportability, robustness, extensibility, etc.). The QA will include overall project-level QA and Software Quality Assurance (SQA) of the CBT system. The SQA consists of various formal and informal reviews, inspections, walkthroughs, measurements and quality audits whose depth and frequency are judiciously tailored to the size, complexity and intended use of the test system and its software.

The QA activities for this project include the following:

1. Formal/Informal Reviews – formal and informal reviews will include the following:
   - Formal reviews will be conducted at the conclusion of each phase of the project and completion of a major task or step in a phase. They are the decision milestones to proceed from one development phase to the next. Entrance and exit criteria for formal reviews will be specified in the QA plan.
   - Informal reviews will be conducted by QA personnel between formal reviews to evaluate progress towards phase completion and/or assess readiness for the formal reviews. Informal SQA reviews also include in-process design and code walkthroughs.
2. Evaluation/Inspections – evaluation and inspections will be conducted periodically by QA to assess conformance to the Project Plan, engineering and software development processes, and contract requirements.
3. Quality Assurance Reporting – monthly status reports will include QA activities performed for the reporting period; results of these activities; problems identified and corrected or action items assigned; status of previous action items; and plans for the next reporting period.

Final Delivery Certification – prior to final delivery of the Screener Readiness Test (SRT), functional and physical configuration audits will be performed on the deliverable system to ensure that the product meets its original requirements and that all changes made through the development process have been properly integrated.

2.2 Project Phases

2.2.1 Phase 1 – Development of Detailed Item And Test Content Requirements

2.2.1.1 Identify Knowledge, Skills and Abilities Expected To Be Acquired During Training

There are a number of sources of information available which can be used to construct a list of Knowledge, Skills, and Abilities (KSA) which can reasonably be expected to be acquired when undergoing the classroom component of initial screener training. These sources include the Screener Personnel Training Guidelines of the ACSSP, the Checkpoint Operations Guide Standard Operating Procedure (SOP) of the ATA, and the FAA's training development guide (Fobes & Neiderman, 1997). The initial KSAs will then be reviewed with Subject Matter Experts (SMEs) to determine that all critical knowledge is adequately represented.

2.2.1.2 Convert The KSAs Into Content Requirements

After the KSAs have been reviewed by SMEs, a tentative test structure and item classification system will be proposed. The test will consist of sub-tests which in turn will be structured by the degree to which particular KSAs are intended to be represented in the item pool. This overall structure will be reviewed by FAA technical center and headquarters personnel and SMEs and will be used to structure the overall item development process.

2.2.1.3 Construct A Test Item Database

ITEM RECORD FIELD	DESCRIPTION
I.D.	Unique item identification number
Type	Textual test or image test
Class	Item's classification using required KSA outline
Readability	Fleisch/Kincaid Readability Measure
Difficulty Index	% tested who passed
Confusability Index	% who chose the most frequently wrong answer
Internal Validity Indices	Item-test correlations or item-subtest correlations
Discrimination Indexes	Training or job performance measures as normalized measures of difference in criterion performance between those who pass and those who fail the item.
Item bias indices	Measures of Differential Item Functioning (DIF)
Item speed statistics	Average and standard deviation of time to complete item

Table 1. Test item database structure.

An item database (refer to Table 1) will be constructed and will consist of individual item records. The general structure of each item record is described in Table 1. These fields are explained in the Item Analysis Strategy section.

2.2.1.4 Perform An Initial Item Analysis

Test material currently available, such as the ATA and Safe Passage training, will be reviewed. Information from these sources which can contribute to an *initial* item analysis will be collected. The current FAA CBT database is one source of such information. A number of analyses are planned using this database.

1) Information to determine item difficulty with a large geographically diverse sample. This analysis will be done to facilitate the item development process by helping to understand what type of questions present difficulties for the target population.
2) Information about how long an item takes to complete. This is very useful in making initial decisions about the probable length of the test or the size of the item pool needed.
3) Estimate parameters for some of the items and units such as test-retest reliability and test homogeneity.
4) Examine item data to determine whether there is evidence of regional bias or differential item functioning associated with specific type of items. Because different airports exhibit quite different ethnic/racial mixes in their workforces, this analysis will provide us with useful information about whether additional effort will be needed to produce a fair test. The relative performance on textual and performance items in this regard will also be determined.

The goal in this initial analysis will be to discover some general characteristics of items that make them good in the sense of reliable, valid, and fair and to use these general principles to guide the item creation process.

2.2.2 Phase II – Development of the Test Prototype Structure and Completion of the Initial Item Pool

2.2.2.1 Develop A Prototype Test Structure

A prototype test structure will be planned including such general characteristics as the existence and number of separate sub-tests and the overall length of the test. All of these decisions will be subject to revision based upon field experience with the test items, but will be used to structure the initial field testing in Phase III.

2.2.2.2 Create An Initial Pool Of Content Items

Using references, such as the Checkpoint Operations Guide and available training materials in current use, human factors engineers experienced in classroom instruction and test construction will develop an initial pool of multiple choice items. At the same time, X-ray images of innocent bags and those containing threats, for use as image items in the performance portion of the test, will be developed.

These items will be given a classification with regard to the KSAs as they are created. The initial test items will be reviewed in consultation with SMEs to determine that the content areas adequately cover the validated KSAs.

2.2.2.3 Create An Initial Pool Of Image Items

Detailed specifications of the types of image items needed will be developed with input from SMEs and FAA personnel. A plan for acquiring the necessary images will then be developed. Most images will be captured using X-ray equipment, baggage, weapons, and simulated Improvised Explosives Devices (IEDs) in the Aviation Security Laboratory. These images will then be downloaded to floppy or zip disk and converted into formats usable by the SRT.

2.2.2.4 Develop the Prototype SRT to Field Test Image Items

Initial programming of the test prototype will be developed sufficiently so that image items can be tested in the next phase. Image items will consist of X-ray images of bags, some of which contain threats including weapons and IEDs. The software at this stage will include the ability to present a series of image items in sequence and to record screener responses.

2.2.2.5 SMEs Review Image Items For Appropriateness And Comprehensiveness

After the initial item set has been created, it will be reviewed with the SMEs to determine whether the item set adequately includes all the major categories of threats. It is intended that screeners will respond to each bag with one of three possible responses: *No Threat, Possible Threat, Definite Threat*. This is a meaningful classification of bags which is included as part of current training and is in use at the checkpoint with different procedures associated with each decision. If possible, a pair of SMEs will review the image sets and determine the appropriate response. Where they cannot agree on a classification, the items will be deleted from the initial set.

2.2.2.6 Document The Process Of Item Development

The prototype test structure (number and composition of sub-tests), and the full set of candidate items, both content and image items, will be included as part of the Content and Image Item Development Report.

2.2.3 Phase 3 – Field Testing of Content and Image Items

2.2.3.1 Preparation For The Field Test

Paper and pencil versions of content items will be constructed and organized into a structure that parallels the proposed test structure. The paper tests will include alternate forms because of the large size of the initial item pool. In cooperation with Security Company Managers at Reagan International Airport (DCA), a large pool (200) of test subjects will be obtained, including experienced screeners who have worked at checkpoints with the threat image projection system. Every effort will be made to maximize the ethnic/gender diversity of this group so that test bias can be assessed according to the principles outlined in the Uniform Federal Guidelines for Selection Tests (1978). If these sample diversity goals can not be met using samples from the Philadelphia International Airport, an additional specific site will be added that will increase the diversity of the sample.

2.2.3.2 Field Test

A field test will be conducted with screeners, who have just completed their initial training, to provide input for item analyses of item difficulty, item-test correlations, and inter-item correlations. Multiple trips (~5) will be taken to DCA during the field test to achieve an adequate sample size. The field test will assess both experienced and new trainees. If a sufficient number of new trainees are not available during the initial testing period, data from experienced screeners will be used to assess concurrent validity, reliability (Kuder-Richardson and test-retest), and test bias/ differential item functioning.

Additionally, pilot testing of the computer test prototype conducted at DCA will particularly target usability testing and the evaluation of reliability and validity of image test items. Every effort will be made to make the conditions of data collection similar to what might be expected with a deployed test.

2.2.3.2 Analysis of Field Test Data

Data collected in the field will be entered into the database and analyses conducted using standard statistical software. These analyses will follow the strategy described in Section 3. The main goal of the analyses at this point will be to identify weakness in the item pool and the test structure and to eliminate items which have poor psychometric qualities.

2.2.3.3 Reporting the Initial Item Validation

Based on the analysis of the initial data, 'bad' items will be identified and eliminated. In addition, the test developers will use the analysis to identify the item characteristics that discriminate bad from good items. Where eliminated items create content deficiencies in the item pool, new items will be constructed using the principles learned in the testing. In this way, a final item set to be used in the prototype will be determined.

Description of the Major Activities

Once the full item pool has been assembled, items will be classified into KSA classes, available item-test, item difficulty calculations carried out, and fairness data assembled. This information will be assembled into a data base and items with bad characteristics will be eliminated.

2.2.4 Phase 4 – Prototype Development

2.2.4.1 Develop Prototype Computer Basis for the SRT

The prototype will have the overall characteristics of the planned final computer platform and will not differ in terms of usability or general appearance of the interface. The following features are planned:

1) Both content- and image-based performance items with responses in a multiple choice format.

2) Individual test sessions with a random selection of items from a larger item database. Test planning and software tracking will guarantee that individual tests are of reasonably uniform

difficulty and have good content sampling and psychometric properties. (This feature will not be fully realized until the final SRT.)

3) Automatic scoring of tests with reporting functions for FAA personnel, guard company, test administrators, etc.
4) Easy to understand instructions and help screens so that the test can be taken with minimum supervision.
5) Security features that make it difficult to download or print test content or test answers and make it difficult to compromise or corrupt the test. (The full set of features may not be realized on the prototype.)
6) Simple to understand security features that make it possible for authorized personnel to alter or add to test content.

2.2.4.2 Develop the Prototype Field Testing Plan

A test plan will be developed to describe the field evaluation and item validation process for the prototype initial screener training test. The test plan shall specify critical operational and technical issues to be resolved by the field test. These issues will be those which have shown themselves to be most critical throughout the period of test development.

2.2.4.3 Field Test of the Prototype

The field test will be of the complete stand alone, computer-based, test prototype with full content and performance test sections. The test subjects will be new screeners who have just completed training. The test and evaluation shall be conducted at a Category X and a Category 1 airport. Where sufficient numbers of new screeners are not available to complete data collection within the time frame planned, experienced screeners will be tested and their data used in analyses where it is appropriate. Prior to the field test, the preliminary User's Guide will be completed to provide support for all field personnel who will use the computer-based test.

2.2.5 Phase 5 – Final Validation and Test Delivery

2.2.5.1 Prepare Final Computer-Based SRT

Based upon the results of the analysis of the final field study, the number and composition of the items on the test will be determined. The final test shall include content and image interpretation items and run on MacIntosh and DOS/PC platforms. This test shall be reliable, valid, non-biased, and secure. The test shall automatically collect and store test performance and individual scores. The test shall be ready for immediate deployment at U.S. airports.

2.2.5.2 Final User's Guide For The Computer-Based SRT

The user's guide in its final form will provide support for all field personnel who will use the computer-based test. The final User's Guide will also contain instructions and step-by-step procedures required.

2.2.5.3 Prepare Final Report

The final report will review and summarize the entire process that was used to assess, develop, and validate the test. By describing the procedures used in test development, it will provide the empirical basis for assessing the reliability, validity, fairness, and security of the test. This report will describe each of the phases on the project and will also include 'lessons learned' to guide future efforts.

3.0 ITEM ANALYSIS STRATEGY

3.1 Preliminary Item Analysis

The item analysis is guided by the end product which is a computerized test of reasonable length which contains both content and performance items. These will be sampled from a larger permanent item pool which is reliable, valid, and fair. Because length of the test is a consideration, it is important to consider elimination of items that have low item-test correlations in order to maintain good reliability.

The first step in the item analysis, content validation and unit construction, is described in Phase 1. The item pool will logically be divided into a small number of subsets, although perhaps no more than the two content and image item sets.

The first item analysis will be performed with data already collected for this population from available item pools. These data will be used to answer a number of general questions about these types of items with this population.

1) Point-biserial item to test/subtest correlations will be determined for current items in order to estimate the mean and variance of these parameters for this population with these types of items. This is a critical statistic because:

    a) Average correlation is a critical determinant of the number of items needed in the test to achieve good reliability.

    b) The variance in the correlation coefficient helps to determine the percentage of new items that will turn out to be unacceptable, so that the initial size of the item pool that is needed can be estimated.

2) Identification of items with poor psychometric properties in the current item sets should make item development more efficient by highlighting general characteristics of these items. Review of these items will also be used to determine principles for eliminating items with poor psychometric properties. Specifically, too hard and too easy items have poor psychometric properties. However, this may be less important than the overall content or construct validity of some such items, in which case they should be retained. This analysis will be used to guide the item creation process as well.

3) If information about ethnic and gender composition of groups or individual screeners is available with these data, it will be possible to determine whether an adverse impact would be associated with current item pools. This information can in turn be used to plan the investigation of DIF.

3.2 Analysis of Field Tested Items

The second item analysis will follow the initial field test of the new pool of items. In the field test, we will collect responses to the pool of candidate items from a large, ethnically diverse number of screeners. Those items will have already been divided into item subsets, as per the overall proposed test structure.

Step 1: Removal of bad items. Based upon a minimum item-test/subset correlation, potentially unacceptable items will be identified. These items will then be closely analyzed for content and distribution of responses to distractors. Bad items will be eliminated or fixed.

Step 2: A corrected Kuder Richardson-20 statistic (Nunnally, 1978) will be used to estimate the reliability of the subtest sampling pool, (r_{NN}) for the N acceptable items in the pool. This process will be repeated for all subtest item pools. These reliability statistics will be used to guide decisions about further changes in the item pools.

The pool reliability statistics need to be used to estimate the reliability of subtests of specific length drawn randomly from the pools. The following formula (Nunnally, 1978) specifies the reliability of a sub-test of *Y* items sampled from a pool of *N* items whose reliability is r_{NN} with *k* as the ratio of *y* to *N*.

$$r_{YY} = \frac{k * r_{NN}}{1 + (k-1) * r_{NN}}$$

An alternate form of the equation shows how to calculate the minimum size of a subtest needed to achieve a specific reliability r_{YY}.

$$k = \frac{r_{YY} * (1 - r_{NN})}{r_{NN} * (1 - r_{YY})}$$

These formulas allow the calculation of reliability of subtests of specific length randomly sampled from the current item pool.

For the overall test, the reliability r_{tt} can be calculated based upon the subtest item pool reliabilities r_{NN}. Specifically:

$$r_{TT} = 1 - \left(\frac{\sum \sigma_i^2 - \sum r_{NN} * \sigma_i^2}{\sigma_s^2} \right)$$

Where σ_I is the standard deviation of subtest i.

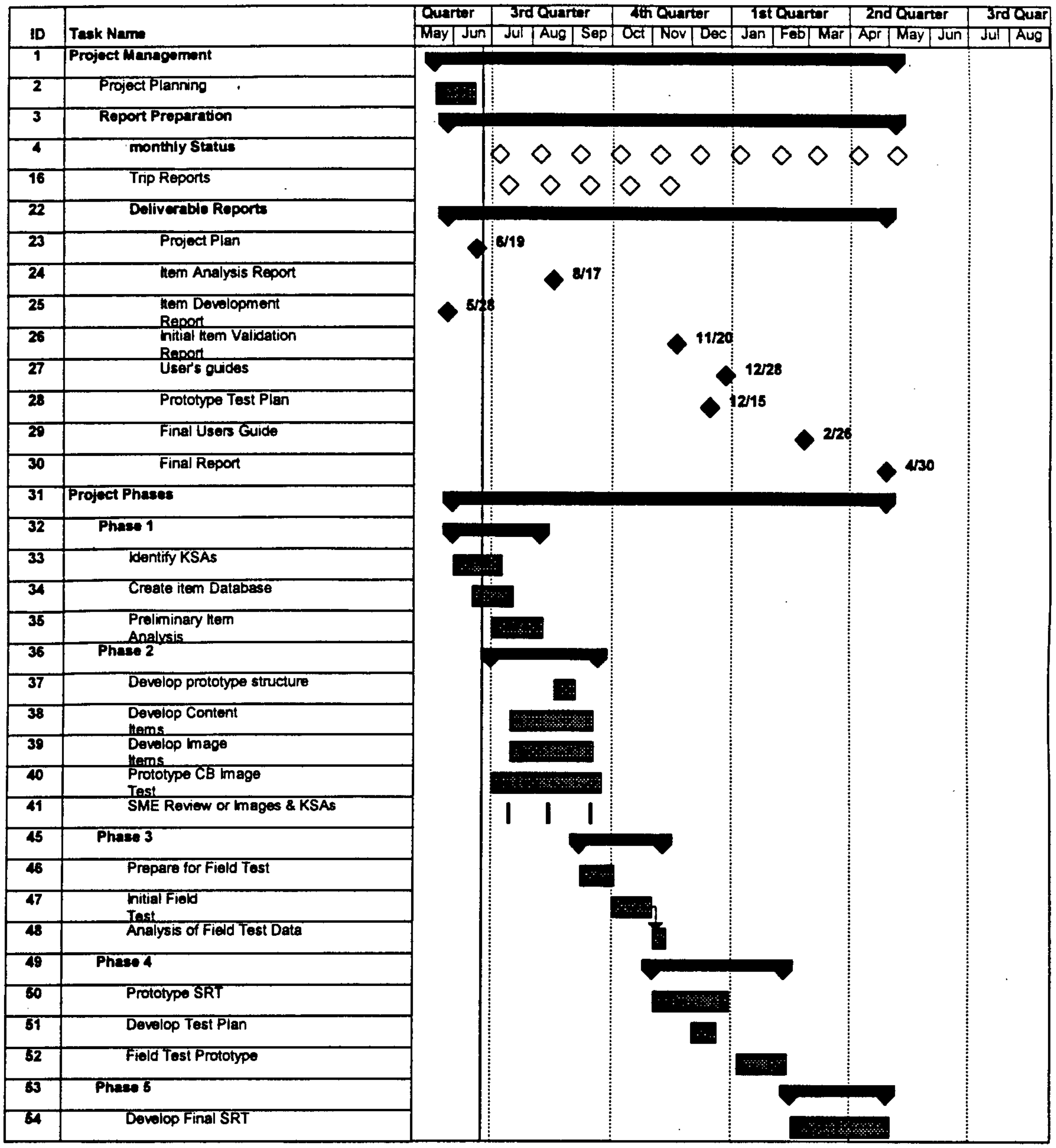
σ_s is the standard deviation of sum of subtests.

The formulas will be used in step 2 to create a test structure using questions sampled from multiple item pools exhibiting good reliability.

Step 3: Analyze overall subtest and test scores, for different racial and gender groups, to identify whether significant differences in test scores exist between different groups and whether an adverse impact may be predicted.

Given an identified risk of adverse impact, compare the distribution of item responses for different racial, gender groups and identify items which exhibit DIF. In this case, DIF is defined as a difference in distribution of item responses among groups matched for ability (defined by total scores or external criterion) but differing in racial/ethnic composition. This approach is more practical than the analysis of DIF using item response theory which requires very large samples of individuals (Hulin, Drasgow, & Parsons, 1983).

4.0 SCHEDULE



ID	Task Name
1	Project Management
2	Project Planning
3	Report Preparation
4	monthly Status
16	Trip Reports
22	Deliverable Reports
23	Project Plan
24	Item Analysis Report
25	Item Development Report
26	Initial Item Validation Report
27	User's guides
28	Prototype Test Plan
29	Final Users Guide
30	Final Report
31	Project Phases
32	Phase 1
33	Identify KSAs
34	Create item Database
35	Preliminary Item Analysis
36	Phase 2
37	Develop prototype structure
38	Develop Content Items
39	Develop Image Items
40	Prototype CB Image Test
41	SME Review or Images & KSAs
45	Phase 3
46	Prepare for Field Test
47	Initial Field Test
48	Analysis of Field Test Data
49	Phase 4
50	Prototype SRT
51	Develop Test Plan
52	Field Test Prototype
53	Phase 5
54	Develop Final SRT

Figure 1. Project Schedule Gantt Chart

5.0 REFERENCES

Fobes, J. L. & Neiderman, E. (1997). *The training development process for aviation screeners*. Technical Report, DOT/FAA/AR-97/46, FAA Technical Center, Atlantic City International Airport, NJ.

Hulin, Drasgow, & Parsons, C. (1983). *Item Response Theory: Application to Psychological Measurement* Dow-Jones, Irwin.

Nunnally, J. C. (1978). *Psychometric Testing*, McGraw Hill N.Y., N.Y.

Uniform Guidelines on Employee Selection Procedure (1978); 43 FR 38295.